Study Comparison K-Means Clustering dengan Algoritma Hierarchical Clustering

Nadya Khalisah Zuhal
Teknik Informatika, Universitas Nusantara PGRI Kediri
E-mail: *nadyadea@gmail.com*

Abstrak – *Study Comparison adalah penelitian dengan cara membandingkan persamaan dan perbedaan sebagai fenomena untuk mencari faktor atau situasi yang menyebabkan perbedaan atau persamaan tersebut. Penelitian ini menggunakan 2 metode clustering yaitu K-Means Clustering dan Algoritma Hierarchical Clustering. Digunakan 2 metode clustering, dengan tujuan untuk mengetahui serta membandingkan metode mana yang lebih bagus dan menghasilkan tingkat kemripian yang optimal. Penelitian ini menggunakan 5 dataset yang berbeda. Dari 5 dataset dibagi menjadi beberapa cluster. Pada hasil penelitian yang telah dilakukan menggunakan jumlah cluster yang optimal dari 2 sampai dengan 8 cluster dengan menggunakan learning rate sebesar 2,02 untuk K-Means Clustering, sedangkan Algoritma Hierarchical Clustering sebesar 0,57 dan nilai centroid acak.*

Kata Kunci *— AHC, K-Means Clustering, Study Comparison*

1. PENDAHULUAN

Keberhasilan suatu sitem dalam mengenali dan mengelompokkan suatu data adalah merupakan suatu tututan. Sebuah pemahaman tentang suatu algoritma pengolahan informasi untuk pembelajaran mesin yang dapat memahami data sebagaimana pemahaman manusia menjadi utama. Beberapa algoritma pengelompokan yang sering digunakan adalah *K-Means Clustering* dan *Algoritma Hierarchical Clustering (AHC)*. Selama ini banyak artikel yang membahas tentang impelementasi dari *K-Means* dan *AHC* akan tetapi banyak pula yang tidak memperhatikan performa yang di dapatkan pada hasil yang diperoleh. Sehingga pada penelitian ini dilakukan sebuah *study comparison* antara *K-Means* dan *AHC*.

Study comparison adalah penelitian dengan cara membandingkan persamaan dan perbedaan sebagai fenomena untuk mencari faktor atau situasi yang menyebabkan perbedaan atau persamaan tersebut. Data sampel yang digunakan dalam metode *K-Means Clustering* dan *Algoritma Hierarchical Clustering (AHC)* yaitu data library.

Pada penelitian ini dilakukan untuk memperbandingkan tingkat kemiripan menggunakan algoritma *K-Means Clustering* dan *Algoritma Hierarchical Clustering (AHC)*. Dan juga menggunakan *Davies Bouldin* Index untuk pengoptimalkan nilai dari 2 metode tersebut. Untuk datasetnya diambil 5 dataset yang berbeda. Dari 5 dataset akan dibagi menjadi beberapa cluster.

2. METODE PENELITIAN

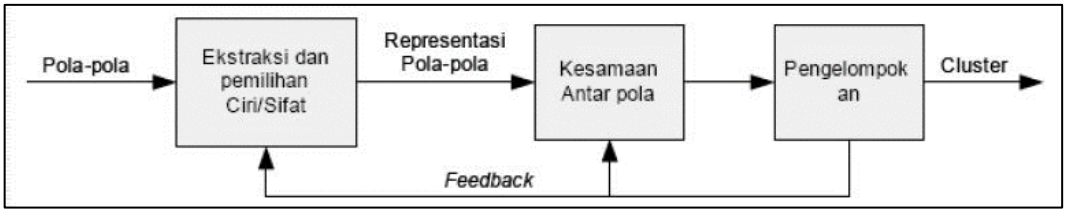
2.1 Study Comparison

Study Comparison adalah penelitian yang dilakukan dengan cara membandingkan persamaan dan perbedaan sebagai fenomena untuk mencari fakta atau situasi yang menyebabkan perbedaan atau persamaan tersebut. Studi ini dimulai dengan mengadakan pengumpulan fakta tentang faktor-faktor yang menyebabkan timbulnya suatu gejala tertentu, kemudian dibandingkan. Setelah mengetahui persamaan dan perbedaan penyebab, selanjutnya ditetapkan bahwa sesuatu faktor yang menyebabkan munculnya suatu gejala pada objek yang diteliti, itulah yang sebenarnya yang menyebabkan munculnya gejala tersebut. Atau dengan memperbandingan faktor atau variabel mana yang paling berpengaruh terhadap perubahan yang terjadi pada hasil penelitian yang sedang dilakukan [2].

2.2 Clustering

Clustering merupakan proses partisi satu set objek data ke dalam himpunan bagian yang disebut dengan *cluster*. Objek yang di dalam *cluster* memiliki kemiripan karaktristik antar satu sama lainnya dan berbeda dengan *cluster* yang lain. Partisi tidak dilakukan secara manual melainkan dengan suatu algoritma *clustering*. Algoritma clustering membagi populasi atau data point dengan sifat yang sama ke beberapa kelompok kecil untuk dikelompokkan. Teknik ini merupakan salah satu algoritma di dalam machine learning yang paling sering digunakan oleh perusahaan untuk melakukan segmentasi kepada customer mereka sehingga dapat meningkatkan

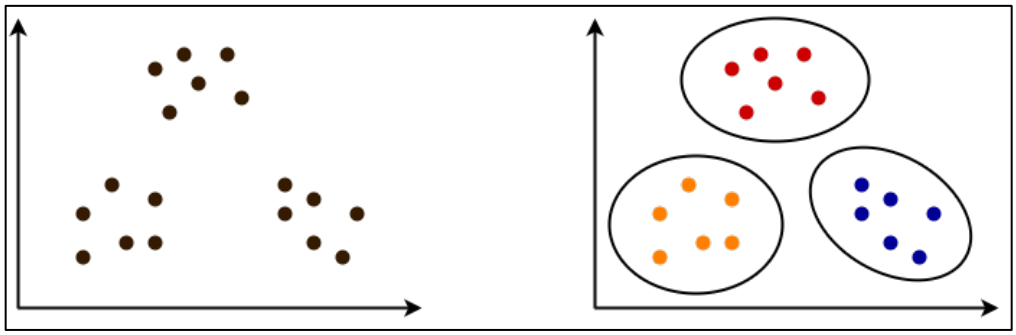
penjualan di perusahaan mereka. Metode *clustering* terdiri dari dua jenis, yaitu *Partitioning clustering* dan *Hierarchical clustering Partitioning clustering* merupakan metode pengelompokan data yang dimulai dengan menentukan terlebih dahulu jumlah *cluster* yang diinginkan (dua *cluster*, tiga *cluster*, atau lain sebagainya). Setelah jumlah *cluster* diketahui, kemudian dilakukan proses *cluster* tanpa mengikuti proses hierarki *Hierachical clustering* adalah suatu metode pengelompokan data yang dimulai dengan mengelompokkan dua atau lebih objek yang memiliki kesamaan paling dekat. Kemudian proses diteruskan ke objek lain yang memiliki kedekatan kedua. Demikian seterusnya sehingga *cluster* akan membentuk seperti pohon, dimana ada hirarki yang jelas antar objek, dari yang paling mirip sampai yang paling tidak mirip [3]. Gambar 1 merupakan gambaran tantang tahapan *clustering*.



Gambar 1. Tahapan *clustering*

2.3 Metode *K-Means Clustering*

K-means clustering merupakan merupakan salah satu metode *cluster analysis* non hirarki yang berusaha untuk mempartisi objek yang ada ke dalam satu atau lebih *cluster* atau kelompok objek berdasarkan karakteristiknya, sehingga objek yang mempunyai karakteristik yang sama dikelompokan dalam satu cluster yang sama dan objek yang mempunyai karakteristik yang berbeda dikelompokan kedalam cluster yang lain. Metode *K-Means Clustering* berusaha mengelompokkan data yang ada ke dalam beberapa kelompok (Gambar 2), dimana data dalam satu kelompok mempunyai karakteristik yang sama satu sama lainnya dan mempunyai karakteristik yang berbeda dengan data yang ada di dalam kelompok yang lain.



Gambar 2. Conto data sebelum dan sesudah dikelompokan dengan *K-Means*.

Dengan kata lain, metode *K-Means Clustering* bertujuan untuk meminimalisasikan objective function yang diset dalam proses *clustering* dengan cara meminimalkan variasi antar data yang ada di dalam suatu cluster dan memaksimalkan variasi dengan data yang ada di *cluster* lainnya juga bertujuan untuk menemukan grup dalam data, dengan jumlah grup yang diwakili oleh variabel K. Variabel K sendiri adalah jumlah *cluster* yang diinginkan. Membagi data menjadi beberapa kelompok. Algoritma ini menerima masukan berupa data tanpa label kelas. Hal ini berbeda dengan *supervised learning* yang menerima masukan berupa vektor *(x1, y1), (x2,y2), ..., (xi,yi),* di mana *xi* merupakan data dari suatu data pelatihan dan *yi* merupakan label kelas untuk *xi* [4].

Algoritma untuk melakukan K-Means *clustering* adalah sebagai berikut:

1. Pilih K buah titik *centroid* secara acak
2. Kelompokkan data sehingga terbentuk K buah *cluster* dengan titik *centroid* dari setiap *cluster* merupakan titik *centroid* yang telah dipilih sebelumnya
3. Perbaharui nilai titik *centroid*
4. Ulangi langkah 2 dan 3 sampai nilai dari titik *centroid* tidak lagi berubah

Proses pengelompokkan data ke dalam suatu *cluster* dapat dilakukan dengan cara menghitung jarak terdekat dari suatu data ke sebuah titik *centroid*. Perhitungan jarak Minkowski dapat digunakan untuk menghitung jarak antar 2 buah data. Rumus untuk menghitung jarak dengan menggunakan persamaan 1:

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \quad \ldots\ldots (1)$$

Di mana:

d_{ij} = jarak antara data i ke data j
x_{ik} = data testing ke-i
x_{jk} = data traning ke-i

Pembaharuan suatu titik centroid dapat dilakukan dengan persamaan 2:

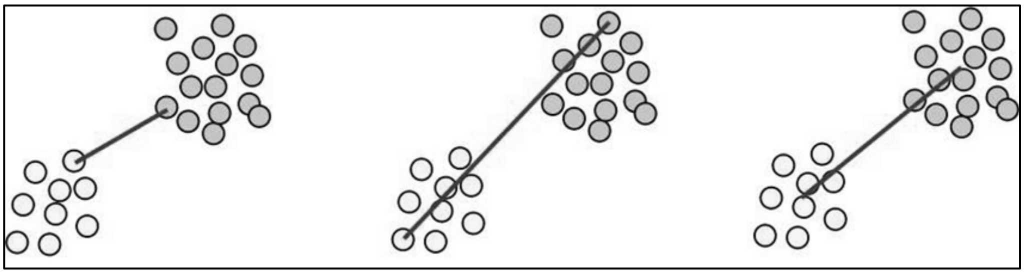
$$\mu_k = \frac{1}{N_k} \sum_{q=1}^{N_k} x_q \quad \ldots\ldots (2)$$

Di mana:

μ_k = titik centroid dari cluster ke-K
N_k = banyaknya data pada cluster ke-K
x_q = data ke-q pada cluster ke-K

2.4 Algoritma Hierarchical Clustering

Algoritma Hierarchical Clustering adalah pengelompokan data dilakukan dengan membuat suatu bagan hirarki (dendrogram) dengan tujuan menunjukkan kemiripan antar data. Setiap data yang mirip akan memiliki hubungan hirarki yang dekat dan membentuk cluster data. Bagan hirarki akan terus berbentuk hingga seluruh data terhubung dalam bagan hirarki tersebut. Cluster dapat dihasilkan dengan memotong bagan hirarki tersebut. Beberapa metode dalam hierarchical clustering yaitu *single linkage, complete linkage, average linkage,* dan *ward's minimum variance.* Gambar 3 menggambarkan perbedaan antara ketiga metode tersebut.



Gambar 3. Perbedaan metode *single linkage, complete linkage, average linkage.*

Secara umum, *hierarchical clustering* dibagi menjadi dua jenis yaitu agglomerative dan divisive. Kedua metode ini dibedakan berdasarkan pendekatan dalam melakukan pengelompokkan data hingga membentuk dendrogram, menggunakan *bottom-up* atau *top-down manner.* Untuk membuat cluster yang memiliki karakteristik yang sama dalam satu anggota cluster yang memiliki karakteristik yang berbeda antar clusternya. Konsep inilah yang mengharuskan proses pembuatan cluster memperhatikan jarak/(dis)similarity/ukuran ketidakmiripan antar data.

Metode penghitungan (dis)similarity yang sering digunakan adalah *Euclidean distance* dan *manhattan distance*, namun bias saja menggunakan pengukuran jarak yang lain, bergantung pada data yang sedang kita analisis. Berikut ini formula dalam perhitungan (dis)similarity tersebut [5]:

1. Euclidean Distance

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (X_{ik} - X_{jk})^2} \quad \ldots\ldots (3)$$

Keterangan
d_{ij} : jarak antara objek *i* dengan *j*

X_{ij} : nilai objek *I* pada variabel ke-*k*
X_{jk} : nilai objek *j* pada variabel ke-*k*
P : banyaknya variabel yang diamati

2. Manhattan Distance

$$d_{xy} = \sum_{i=1}^{n} |(x_i - y_i)| \quad (4)$$

2.5 Dataset

Dataset adalah sebuah kumpulan data yang berasal dari informasi-informasi pada masa lalu dan siap untuk dikelola menjadi sebuah informasi baru. Data yang digunakan dalam penelitian ini ada 5 data eksperimental yang diambil dari UCI Machine Learning Repository. Adapun rincian dari 5 dataset yang digunakan, yaitu:

1. Data1: BuddyMove Dataset
   Datanya diambil dari https://archive.ics.uci.edu/ml/datasets/BuddyMove+Data+Set#, dataset tersebut terdapat 249 data latih. Data ini berisi informasi minat pengguna yang diekstrak dari tinjauan pengguna yang diterbitkan di holidayiq.com tentang beragam jenis kepentingan di India Selatan.
2. Data 2: Heart Disease Cleveland UCI Dataset
   Datanya diambil dari https://archive.ics.uci.edu/ml/datasets/Heart+Disease, dataset tersebut terdapat 297 data latih. Database ini berisi 76 atribut, tapi semua percobaan yang diterbitkan merajuk untuk menggunakan bagian dari 14 dari mereka. Khususnya, database Cleveland adalah satu-satunya yang digunakan oleh para peneliti ML.
3. Data 3: Cryotherapy Dataset
   Datanya diambil dari https://archive.ics.uci.edu/ml/datasets/Cryotherapy+Dataset+, dataset tersebut terdapat 90 data latih. Data ini berisi informasi tentang hasil perawatan kuntil yang menggunakan krioterapi.
4. Data 4: Travel Reviews Dataset
   Datanya diambil dari https://archive.ics.uci.edu/ml/datasets/Travel+Reviews, dataset tersebut terdapat 981 data latih. Data ini berisi ulasan tentang destinasi dalam 10 kategori yang disebutkan di seluruh Asia Timur. Setiap kategori adalah dipetakan sebagai yang sangat baik(4), sangat baik(3), rata-rata(2), miskin(1) dan mengerikan(0) dan peringkat rata-rata.
5. Data 5: Wine Dataset
   Datanya diambil dari https://archive.ics.uci.edu/ml/datasets/wine, dataset tersebut terdapat 178 data latih. Data ini berisi penggunaan analisis kimia menentukan asal mula anggur.

2.6 Evaluasi

Evaluasi *clustering* dilakukan dengan tujuan untuk mengetahui seberapa baik kualitas dari hasil *clustering*. Pada tahap ini dilakukan perbandingan antara dua algoritma untuk didapatkan hasil penelitian dengan pengukuran tingkat kemiripan berupa nilai angka. Untuk perhitungannya menggunakan *Davies Bouldin Index(DBI)*.

2.7 Davies Bouldin Index (DBI)

Davies Bouldin Index (DBI) diperkenalkan oleh David L. Davies dan Donald W. Bouldin pada tahun 1979 adalah metric untuk mengevaluasi hasil algoritma clustering Evaluasi menggunakan *DaviesBouldin Index* ini memiliki skema evaluasi internal *cluster*, dimana baik atau tidaknya hasil *cluster* dilihat dari kuantitas dan kedekatan antar data hasil *cluster Davies-Bouldin Index* merupakan salah satu metode yang digunakan untuk mengukur validitas *cluster* pada suatu metode pengelompokan, kohesi didefinisikan sebagai jumlah dari kedekatan data terhadap titik pusat *cluster* dari *cluster* yang diikuti. Sedangkan separasi didasarkan pada jarak antar titik pusat *cluster* terhadap *cluster*nya. Pengukuran dengan *Davies-Bouldin Index* ini memaksimalkan jarak inter-*cluster* antara *cluster Ci* dan *Cj* dan pada waktu yang sama mencoba untuk meminimalkan jarak antar titik dalam sebuah *cluster*. Jika jarak inter-*cluster* maksimal, berarti kesamaan karakteristik antar-masing-masing *cluster* sedikit sehingga perbedaan antar-*cluster* terlihat lebih jelas. Jika jarak intra-*cluster* minimal berarti masing-masing objek dalam *cluster* tersebut memiliki tingkat kesamaan karakteristik yang tinggi [6]. Rumus untuk menghitung *Davies- Bouldin Index* (BDI) dengan persamaan 5:

$$DBI = \frac{1}{k} . \sum_{i=1}^{k} R_i \quad (5)$$

Dengan

$$R_i = \max R_{ij} \quad \text{(6)}$$

Dan

$$R_{ij} = \frac{var(C_i)+var(C_j)}{\|C_i - C_j\|} \quad \text{(7)}$$

Dimana:

C_i : *cluster* I dan c_i adalah centroid dari *cluster* i.

3. HASIL DAN PEMBAHASAN

3.1 Hasil Pengujian Dan Pembahasan

Pada penelitian ini, untuk pengujian dilakukan dengan cara pengambilan 5 dataset. Data tersebut dilakukan untuk menentukan satu persatu apakah memiliki kemiripan dengan menggunakan 2 metode algoritma.

Untuk mengetahui efektifitas dari metode yang diusulkan yaitu penentuan titik pusat *cluster* berdasarkan *Purity minimum-maximum* dengan metode penentuan titik pusat *cluster* awal pada algoritma K-Means secara konvensional terhadap proses *clustering*, maka dilakukan perbandingan total rata-rata hasil evaluasi *clustering* dari keempat *dataset*. Perbandingan rata-rata hasil evaluasi *clustering* dari kedua metode terhadap keempat *dataset* yang digunakan dapat dilihat pada tabel 1. Dari hasil pengujian, tingkat kemiripan yang didapat dari 5 dataset dengan algoritma K-Means sebesar 2,02, sedangkan algoritma AHC sebesar 0,57. Tabel 1 menunjukan rincian hasil pengujian yang telah dilakukan.

Tabel 1. Hasil Pengujian

	Jumlah Data	*DBI*			
		K Means	*Jumlah Cluster optimal K-Means*	*AHC*	*Jumlah Cluster optimal AHC*
Data 1	249	1.15	8 cluster	0.61	
Data 2	297	4.61	2 cluster	0.68	8 cluster
Data 3	90	1.48	3 cluster	0.33	2 cluster
Data 4	981	1.01	8 cluster	0.78	2 cluster
Data 5	178	1.85	2 cluster	0.45	8 cluster
Jumlah		10.1		2.85	
Rata-rata		2.02		0.57	

4 SIMPULAN

Setelah melalui tahap pengujian tingkat kemiripan pada 2 metode *algoritma K-Means Clustering* dan *Algoritma Hierarchical Clustering* dengan menggunakan perhitungan *Davies Bouldin Index*, maka dapat diambil kesimpulan antara lain:

1. Hasil *cluster* dipengaruhi dari nilai titik pusat *cluster(centroid)* dan jumlah data yang digunakan. Selain itu perbedaan pengambilan data pusat *cluster* awal yang digunakan juga akan mempengaruhi hasil akhir dari pengoptimalan untuk 2 metode *clustering*.
2. Dari hasil pengujian tingkat kemiripan yang telah dilakukan untuk metode *algoritma K-Means Clustering* memiliki tingkat kemiripan sebesar 2,02. Sedangkan pada metode *Algoritma Hierarchical Clustering* memiliki tingkat kemiripan sebesar 0.57. Hal ini disebabkan 2 metode *clustering* dan *Davies Bouldin Index* akan menentukan jumlah cluster yang paling optimal berdasarkan kemiripan dari 5 dataset tersebut.

3. Penggunaan *Davies Bouldin Index* (DBI) menghasilkan *cluster* set yang paling optimal.

5 SARAN

Pada penelitian ini adalah memaksimalkan lagi tingkat kemiripan lebih dari hasil pengujian yang dilakukan dan datasetnya berbeda. Untuk menggunakan metode yang lain dan dibandingkan agar lebih optimum. Kombinasi metode penentuan titik pusat *Purity* dengan metode yang lain untuk menentukan dan memilih titik pusat *cluster* awal yang lebih baik untuk *dataset* yang sama ataupun *dataset* yang berbeda. Penerapan metode evaluasi *clustering* yang lain pada *dataset* yang memiliki jumlah data lebih besar untuk mengevaluasi hasil yang lebih baik terhadap hasil *clustering*.

DAFTAR PUSTAKA

[1] Irhamni, Firli, Fitri Damayanti, Bain Khusnul K., Miffachul A. 2014. Optimalisasi Pengelompokan Kecamatan Berdasarkan Indikator Pendidikan Menggunakan Metode Clustering Dan Davies Bouldin Index. Diambil 20 Januari 2022 dari https://stmiksznw.ac.id/jurnal/index.php/teknimedia/article/view/27

[2] Kresna. 2019. Pengertian Penelitian Comparative study (studi perbandingan) (skripsi dan tesis). Diambil 10 Januari 2022 dari http://konsultasiskripsi.com/2019/02/14/pengertian-penelitian-comparative-study-studi-perbandingan-skripsi-dan-tesis/

[3] Edy, Irwansyah. 2017. *CLUSTERING*. Di ambil 20 Desember 2021 dari https://socs.binus.ac.id/2017/03/09/clustering/

[4] Syafnidawaty. 2020. *K-means Clustering*. Di ambil 20 Desember 2021 dari https://raharja.ac.id/2020/04/19/k-means-clustering/

[5] S, Inayatus, Nabiilah A.F. 2021. Introduction To Hierarchical Clustering. Di ambil 12 Januari 2022 dari https://algotech.netlify.app/blog/introduction-to-hierarchical-clustering/

[6] Davies, D. L.; Bouldin, D. W. "A Cluster Separation Measure", IEE Transactions on Pattern Analysis and Machine Intelligence (2):224, 1979.